

## DEVELOPMENT OF NEW COMPUTATIONAL AMINO ACID PARAMETERS FOR PROTEIN STRUCTURE/FUNCTION ANALYSIS WITHIN THE RESONANT RECOGNITION MODEL

Elena Pirogova<sup>1</sup>, Irena Cosic<sup>1</sup>, member IEEE

<sup>1</sup>Bioelectronics Group, Department of Electrical and Computer Systems Engineering,  
Monash University, Melbourne, Australia

e-mail: elena.pirogova@eng.monash.edu.au

e-mail: irena.cosic@eng.monash.edu.au

**Abstract**—The Resonant Recognition Model (RRM) is a physico-mathematical model developed for analysis of protein and DNA sequences. Biological function of proteins and their 3D structures are determined by the linear sequences of amino acids. Previously, the electron-ion interaction potentials (EIIP) of amino acids have been used to determine the characteristic patterns of different proteins independent of their biological activity. In this study, the effect of various other amino acid parameters on periodicity, obtained using the RRM, were assessed. Here, we are proposing new computational amino acid parameters that could be used successfully for protein analysis instead of EIIP within the RRM.

### I. INTRODUCTION

The *Resonant Recognition Model* is based on the finding that there is a significant correlation between spectra of the numerical presentation of amino acid and their biological activity [1,2]. With the rapid expansion of protein databases, the biological function of newly sequenced proteins and determination of their relationship with defined functional families has become a real problem. Consequently, the *de novo* design of protein analogues with the desired biological activity is the ultimate goal of these studies. The RRM is a model that interprets the protein primary sequence [1,2]. The RRM has been employed in this study to investigate the effect of various amino acid parameters on the determination of biological profile of the protein groups under examination.

### II. METHODOLOGY

The Resonant Recognition Model (RRM) interprets the protein linear information using signal analysis methods [1,2]. It has been shown that certain periodicities (frequencies) within the distribution of energies of delocalised electrons along the protein are critical for protein biological function (i.e. interaction with its target).

The RRM, used in this study, involves transformation of the amino acid sequence into a numerical sequence and then analysis of this sequence by appropriate digital signal processing methods (FFT, wavelets etc.). To determine the common frequency components in the spectra for a group of proteins, the multiple cross-spectral function was used. Peaks

in this function denote common frequency components for the sequences analysed. Through an extensive study, the RRM has reached a fundamental conclusion: *one RRM characteristic frequency characterizes one particular biological function or interaction* [2,3].

Once the RRM characteristic frequency for a particular biological function or interaction has been determined, it is possible to identify then the individual amino acids so called “hot spots”, or domains that contribute mostly to the characteristic frequency and thus to protein’s biological function as well [3,4].

The RRM is based on the representation of the protein primary structure as a numerical series by assigning to each amino acid a physical parameter value relevant to the protein’s biological activity. A number of amino acid indices (402 have been published up to now [9]) have been found to correlate in some way with the biological activity of the whole protein. Previous investigations [1-5] have shown that the best correlation can be achieved with parameters, which are related to the energy of delocalised electrons of each amino acid. These findings can be explained by the fact that the electrons delocalised from the particular amino acid have the strongest impact on the electronic distribution of the whole protein. In previous studies [1-3], the energy of delocalised electrons (calculated as the electron-ion interaction pseudopotential, EIIP [5]) of each amino acid residue was employed. The resulting numerical series then represented the distribution of the free electrons’ energies along the protein. The values of EIIP for each amino acid were mathematically obtained from an approximate pseudopotential model [5]. This numerical series was then converted into a discrete Fourier spectrum, which carried the same information content about the arrangement of amino acids in the sequence as the original numerical sequence [1,2].

Although the EIIP was used successfully in our previous studies, recently we have shown that similar, consistent results could be obtained by using Ionisation Constant of amino acid (IC) parameter value instead of the EIIP to represent each amino acid in the sequence [6-8]. IC is a measurable physical parameter and is more exact in comparison with EIIP, which was calculated and includes a lot of approximations. Furthermore, we have analysed also different amino acid properties presented in AAIndex

## Report Documentation Page

<b>Report Date</b> 25 Oct 2001	<b>Report Type</b> N/A	<b>Dates Covered (from... to)</b> -
<b>Title and Subtitle</b> Development of New Computational Amino Acid Parameters for Protein Structure/Function Analysis Within the Resonant Recognition Model		<b>Contract Number</b>
		<b>Grant Number</b>
		<b>Program Element Number</b>
<b>Author(s)</b>	<b>Project Number</b>	
	<b>Task Number</b>	
	<b>Work Unit Number</b>	
<b>Performing Organization Name(s) and Address(es)</b> Bioelectronics Group Department of Electrical and Computer Systems Engineering Monash University Melbourne, Australia		<b>Performing Organization Report Number</b>
<b>Sponsoring/Monitoring Agency Name(s) and Address(es)</b> US Army Research Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500		<b>Sponsor/Monitor's Acronym(s)</b>
		<b>Sponsor/Monitor's Report Number(s)</b>
<b>Distribution/Availability Statement</b> Approved for public release, distribution unlimited		
<b>Supplementary Notes</b> Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.		
<b>Abstract</b>		
<b>Subject Terms</b>		
<b>Report Classification</b> unclassified	<b>Classification of this page</b> unclassified	
<b>Classification of Abstract</b> unclassified	<b>Limitation of Abstract</b> UU	
<b>Number of Pages</b> 4		

Database [9]. Each of the 20 amino acids has multifaceted properties that are responsible for the specificity and diversity of protein structure and function. A large body of experimental and theoretical research has been performed to characterize different kinds of properties of individual amino acids and to represent them in terms of the numerical index. The Cluster analysis of amino acid indexes method finds the best sets of parameters that discriminate different groups of sequence data [9]. One of the important factors in this analysis is how the amino acid sequence should be represented. The amino acid sequence may be considered as a sequence of numerical values reflecting various aspects of amino acid residues, such as: hydrophobicity and bulkiness. An Amino Acid Index (K. Tomii and M. Kanehisa et al., 2000) is a set of 20 numerical values representing any of the different physicochemical and biochemical properties of amino acids.

Once the database was established, correlation calculations were carried out (using Microsoft Excel program) sequentially between each of the database's parameters and the referential EIIP values in order to test for any linear relationship between them. Parameters having correlation coefficients in excess of  $\pm 0.5$  were deemed to be the most strongly correlated with EIIP and were thus isolated [11]. Those selected parameters were: P001- $\alpha$ -CH chemical shifts, H085-Localised electrical effect and H371- Normalized frequency of chain reversal D [9].

Sequences from six protein functional groups were analyzed. Each amino acid in the sequence was represented by corresponding value of selected parameters instead by EIIP value as it was done previously. Numerical series obtained in this way were analyzed and a multiple cross-spectral analysis was performed for each protein group. Results obtained revealed that P001, H085 and H371 amino acid parameters can be used for protein structure/function analysis within the RRM, so far as they satisfied all RRM criteria. However, those three parameters were not equally efficient accordingly to correlation coefficient and S/N values. It is important to mention that the correlation coefficients values of studied parameters P001, H085 and H371 are smaller than the corresponding correlation coefficient values of IC with EIIP. Based on our previous investigations we could conclude that the significance of correlation between selected parameters is then reflected into the analogy of consensus spectra of analyzed proteins and the similarity of characteristic frequencies. Thus, the correlation coefficient value of studied parameter could be considered as one of the major factors influenced on the selection of the analyzed parameter for its further usage for protein analysis within the RRM. Finally, after comparison obtained results we have concluded that selected parameters P001, H085 and H371 wouldn't be solely the best parameters to use in further protein modeling within the RRM.

Therefore, in this study we have tried different mathematical combinations of selected amino acid parameters (IC, P001, H085 and H371) with the aim to find among them the most correlated with EIIP parameter. This new parameter should

be more significantly correlated with EIIP than selected above parameters.

As a result of our calculations, we are proposing here new computational parameter EEIC. This parameter is a simple mathematical combination of two selected previously parameters: IC (Ionisation constant of amino acid) and H085 (Localised electrical effect). We have chosen this parameter because of his strong correlation with the EIIP (correlation coefficient is -0.807).

Here, we have analysed and compared the following amino acid parameters: EIIP, IC, EE (H085 in previous study, renamed for further convinience) and computational parameter ICEE (ICEE=IC-EE). The selected parameter values are presented in Table 1.

TABLE 1.  
Analysed parameter values and their correlation coefficients with EIIP.

Amino Acid	Parameter			
	EIIP	IC	EE	ICEE
<b>L</b>	0	2.40	-0.01	2.41
<b>I</b>	0	2.40	-0.01	2.41
<b>N</b>	0.0036	2.20	0.06	2.14
<b>G</b>	0.0050	2.46	0.00	2.46
<b>V</b>	0.0057	2.35	0.01	2.34
<b>E</b>	0.0058	2.30	0.05	2.23
<b>P</b>	0.0198	2.00	0.00	2.00
<b>H</b>	0.0242	2.30	0.08	2.22
<b>K</b>	0.0371	2.20	0.00	2.20
<b>A</b>	0.0373	2.30	-0.01	2.31
<b>Y</b>	0.0516	2.20	0.03	2.17
<b>W</b>	0.0548	2.37	0.00	2.37
<b>Q</b>	0.0761	2.06	0.07	2.01
<b>M</b>	0.0823	2.17	0.04	2.13
<b>S</b>	0.0829	2.10	0.11	1.99
<b>C</b>	0.0829	1.96	0.12	1.84
<b>T</b>	0.0941	2.09	0.04	2.05
<b>F</b>	0.0946	1.98	0.03	1.95
<b>R</b>	0.0959	1.82	0.04	1.78
<b>D</b>	0.1263	1.88	0.15	1.73
<b>Cor. Coef.</b>		<b>-0.794</b>	<b>0.564</b>	<b>-0.807</b>

### III. RESULTS.

Sequences form different functional groups (glucagon, lysozyme, hemoglobin, cytochrome C, EGF, TGF, PDG, TNF, interleukin, myoglobin, viral oncogen, proto-oncogene, oncogene, and p53 protein) were investigated.

A multiple cross-spectral analysis was performed for each protein group using selected parameter values. As a result protein characteristic frequencies for each protein functional group were obtained.

The signal-to-noise value for the characteristic frequency was calculated as the ratio between signal intensity at the particular peak frequency and the spectrum mean value.

The peak frequency and S/N values for each protein group are shown in Table 2.

#### IV. DISCUSSION

The assumption is that each specific biological function is characterised by a single frequency. Results obtained have shown that all selected for analysis parameters (IC, EE and ICEE) generate in consensus spectrum one dominant peak corresponding to common biological activity of selected proteins and are performed accordingly to criteria determined within the RRM.

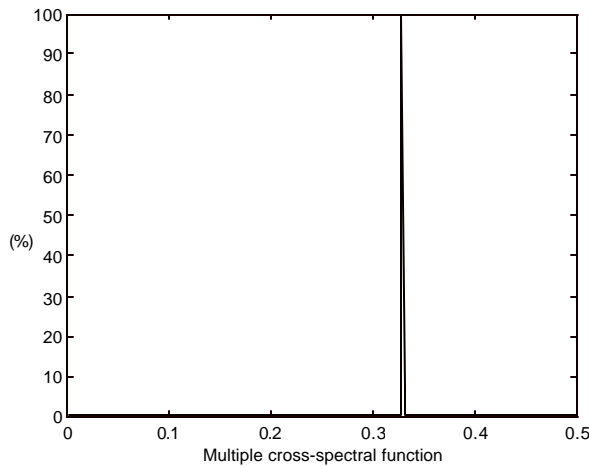


Fig. 1. Multiple cross-spectral function of Lysozyme using EIIP parameter.

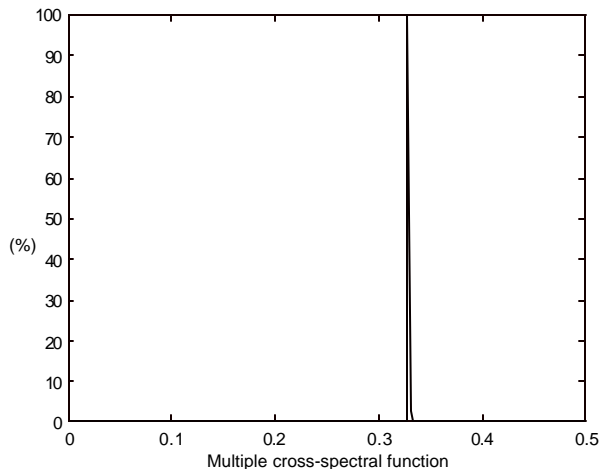


Fig. 2. Multiple cross-spectral function of Lysozyme using IC parameter.

It could be observed from the Table 2. For Lysozyme, Hemoglobin, Cytochrome C, EGF, TGF, PDG, TNF, Interleukin, Proto-oncogene, Oncogene and P53 proteins the same characteristic frequencies were obtained using parameters IC and ICEE. This similarity is very expectable, as both parameters are strongly correlated (0.9837). Their high correlation is reflected then in the analogy of consensus spectrum.

It should be noted that for Lysozyme, Hemoglobin, Cytochrome C, EGF and TGF the same characteristic

frequencies were obtained using parameters EIIP, IC and ICEE.

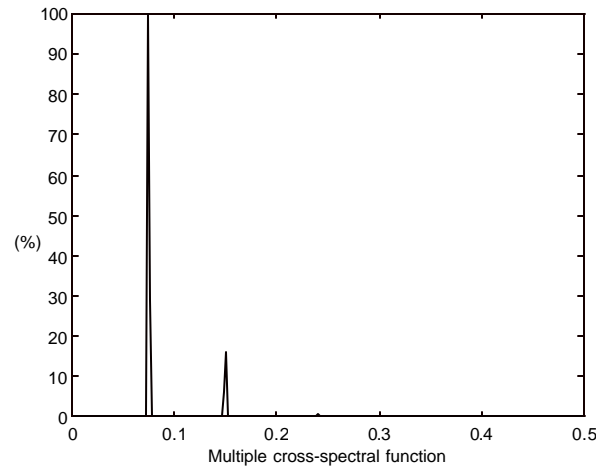


Fig. 3. Multiple cross-spectral function of Lysozyme using EE parameter.

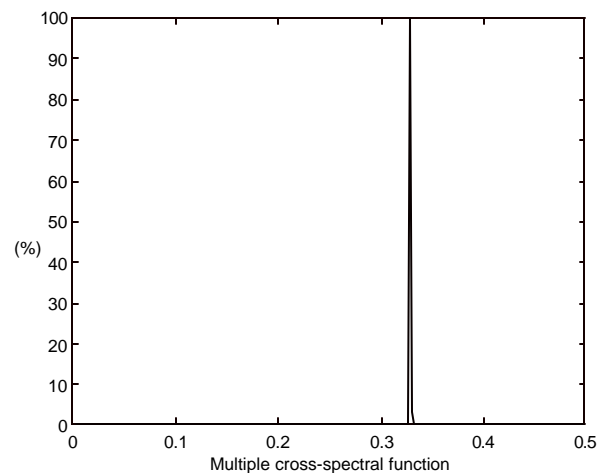


Fig. 4. Multiple cross-spectral function of Lysozyme using ICEE parameter.

This fact could be explain also by the significant correlation between these parameters. The analogy of the peak frequencies implies that in this particular frequency the analyzed protein group reveals the same specific biological function. Of interest is that using all analysed parameters EIIP, IC, EE and ICEE we could gain the same characteristic frequencies only for two protein groups: EGF and TGF. Furthermore, for Viral oncogene we obtained the similarity of frequencies using parameters EIIP and EE.

It could be also observed from the Table 2. that characteristic frequencies are different for the other protein groups. It could be explained by the fact that frequencies are different for different biological functions. Thus, each of analyzed parameters EIIP, IC, EE and ICEE allows us to detect a single frequency or frequencies, which are relevant to the specific biological function(s) of the studied protein sequences.

## V. CONCLUSION

The aim of this study was to test the possible usage of proposed parameters IC, EE and ICEE instead of the EIIP for structure/function analysis of different proteins within the RRM. Results obtained reveal that among selected and comparable amino acid parameters only IC and ICEE are currently the most appropriate parameters for RRM analysis of different unrelated protein families. We can conclude this as far as these two parameters satisfy all RRM criteria, generate one prominent peak corresponding for biological activity of whole protein group and are highly correlated with EIIP. Despite it is still difficult finally to conclude which parameter would be solely the best parameter to use in the RRM. The selectivity of protein interactions within the amino acid sequence could be identified if appropriate physical parameters are used. Thus, more research needs to be done in this direction. Therefore, we plan to measure for the first time the dielectric properties of twenty single amino acids and investigate the possible usage of these new values in the RRM. These newly measured parameters will replace then the EIIP values, which were mathematically calculated from an approximate pseudopotential model [7].

## References

- [1]. Cosic, I. (1994) "Macromolecular bioactivity: Is it Resonant Interaction between Molecules? - Theory and Applications", IEEE Trans. On BME, 41, 1101-1114.
- [2]. Cosic, I. (1997) The Resonant Recognition Model of Macromolecular Bioactivity.
- [3]. Cosic, I., Drummond, A.E., Underwood, J.R. and Hearn, M.T.W. (1994) In vitro inhibition of the actions of basic FGF by a novel 16 amino acid peptide, Molecular And Cellular Biochemistry, 130,1-9.
- [4]. Krsmanovic, V., Biquard, J. -M., Sikorska-Walker, M., Cosic, I., Desgranges, C., Trabaud, M. -A., Whitfield, J.F., Durkin, J.P., Achor, A., Hearn, M.T.W. (1998) Investigations into the cross-reactivity of rabbit antibodies raised against nonhomologous pairs of synthetic peptides derived from HIV-1 gp120 proteins.
- [5]. Veljkovic, I. And Slavic, M. (1972) General Model of Pseudopotentials, Physical Review Let. 29, 105-108.
- [6]. I.Cosic, E.Pirogova, (1998) "Applications of Ionization Constant of Amino Acids for Protein Signal Analysis Within the Resonant Recognition Model", Proceedings of 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Hong Kong, October, Vol.20, No.2, 1072-1075.
- [7]. Cosic, I., Fang, Q., Pirogova, E., "Modification of the RRM model using wavelet transform and Ionisation constant to predict protein active sites", IEEE EMBS, 21, Vol.2, 1215-1217.
- [8]. Cosic, I., Pirogova, E., "Usage of Ionization constant of amino acids for protein signal analysis within the RRM - Application to Oncogene", IEEE-EMBS Asia-Pacific Conference on Biomedical Engineering (full paper, accepted for publications).
- [9]. Kawashima, S. and Kanehisa, M. (2000) Aaindex: amino acid index database, Nucleic Acid Res., 28, 374.
- [10]. Pirogova, E., Cosic, I., (2001) "Examination of amino acid indices within the Resonant Recognition Model", Proc. of the 2nd Conference of the Victorian Chapter of the IEEE EMBS, 124-127.
- [11]. Oyster, C.K., Hanten, W>P. and Liorence, L.A. (1987) Intro. To Research: A Guide for the Health Science Professional, 170-176.

TABLE 2.  
Peak frequency and signal-to-noise values for protein groups.

Protein group	Parameter							
	EIIP		IC		EE		ICEE	
	Freq	S/N	Freq.	S/N	Freq.	S/N	Freq.	S/N
Glucagon	0.0879	122.0	0.0859	94.0	0.1523	140.1	0.3340	91.8
Lysozyme	0.3281	238.0	0.3281	249.0	0.0742	167.3	0.3281	248.8
Hemoglobin	0.0254	256.0	0.0254	256.0	0.4102	249.1	0.0254	256.0
Cytochrome C	0.4746	247.0	0.4746	209.0	0.4219	232.3	0.4746	255.4
EGF	0.0605	246.0	0.0605	256.0	0.0684	171.2	0.0605	255.6
TGF	0.0137	29.0	0.0117	38.0	0.0117	42.7	0.0117	54.2
PDG	0.4199	75.0	0.1934	32.0	0.1797	21.5	0.1934	29.8
TNF	0.0527	119.0	0.0762	80.0	0.1445	38.5	0.0762	65.0
Interleukin	0.0410	191.0	0.0957	192.0	0.1016	236.5	0.0957	182.5
Myoglobi n	0.2539	255.4	0.0137	191.3	0.1270	256.0	0.4824	186.5
Viral oncogene	0.1611	375.0	0.4844	368.0	0.1611	468.3	0.3115	236.5
Proto-oncogene	0.0576	219.0	0.4189	260.0	0.1641	222.7	0.4189	196.0
Oncogene	0.0332	505.0	0.4180	512.0	0.1641	347.6	0.4180	458.9
P53	0.1943	156.0	0.1494	348.0	0.0381	289.1	0.1494	401.2